Comput. Math. Applic. Vol. 16, No. 9, pp. 737–745, 1988
Printed in Great Britain.

0097-4943/88 $3.00 + 0.00

A CONJUGATE DIRECTION GRADIENT METHOD WITH RECONNAISSANCE STEPS FOR UNCONSTRAINED MINIMIZATION

J. A. Snyman† and M. van Rooyen
Department of Mathematics and Applied Mathematics, University of Pretoria, Pretoria, Republic of South Africa

(*Received* 3 *March* 1988)

Communicated by E. Y. Rodin

Abstract—This paper presents a new conjugate direction, and thus quadratic terminating, method for unconstrained minimization. The algorithm consist of a basic cycle which is repeated iteratively until convergence is achieved. In each cycle two phases are executed. In the first reconnaissance points are generated in the direction of descent by using gradient information. A novel feature is that in the second phase, in which the line searches are performed, the searches are initiated from points divorced from the current point with the first line search starting from the furthest reconnaissance point. An analysis of the convergence properties of the method is performed and consideration is given to the practical problem of the economic generation of reconnaissance points. The working of the method is demonstrated by application to the Rosenbrock test function.

1. INTRODUCTION

We consider the problem of minimizing a continuously differentiable function $f(\mathbf{x})$, $\mathbf{x} \in R^n$; where at any point $\mathbf{x}$ the gradient vector $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ can be calculated but no information regarding the second derivatives is explicitly given.

Many descent methods using gradient information are available that can solve the above problem by generating a sequence of points $\mathbf{x}^i$, $i = 0, 1, 2, \ldots$, that converges to a local minimizer $\mathbf{x}^*$. Of these methods the simplest is the method of steepest descent while the conjugate gradient method of Fletcher and Reeves [1] is probably the most well-known of the more sophisticated gradient techniques. The quasi-Newton methods that attempt to approximate the Hessian matrix of second derivatives may also be considered to fall in the above general class, in that they also explicitly use only gradient information in the calculation of search directions. Many of these methods possess the property of quadratic termination since they lead to the optimum of a positive definite quadratic function in at most n exact line searches. However, the most general feature of all these methods is that in calculating the next search direction at the current point $\mathbf{x}^i$ they only use local information regarding the gradient, i.e. $\mathbf{g}^i = \mathbf{g}(\mathbf{x}^i)$, and information supplied by *preceding points*.

In contrast to the above established approach we propose here a gradient method that cyclically computes the next n search directions by first computing an additional $n - 1$ *reconnaissance points* "ahead" of the current point. In doing so we hope to utilize additional global information regarding the behaviour of the function in the direction in which progress is being made. We show that this method also possesses the desired property of quadratic termination and prove some convergence results for the more general case. Although the principal purpose of this paper is the presentation and analysis of this new approach, rather than an exhaustive experimental study of the performance of the new method relative to other more established and refined methods, we do give the computed descent path for the well-known Rosenbrock problem to illustrate the behaviour of the reconnaissance method.

†To whom all correspondence should be addressed.

2. THE BASIC STRUCTURE OF A RECONNAISSANCE METHOD

Consider the basic structure of a line search algorithm using gradient information. Given an initial estimate $\mathbf{x}^0$ the ith interation, $i = 1,2, \ldots$, consists of the following steps:

(a) determine the search direction $\mathbf{s}^i$;
(b) find λ_i to minimize $f(\mathbf{x}^{i-1} + \lambda \mathbf{s}^i)$ with respect to λ;
(c) $\mathbf{x}^i \leftarrow \mathbf{x}^{i-1} + \lambda_i \mathbf{s}^i$.

In Step (a) $\mathbf{s}^i$ is computed using local gradient information plus, usually, additional information relating to previous points more distant from $\mathbf{x}^*$. The question now arises as to whether advantage could be gained by taking reconnaissance jumps "ahead" of the current point. With this further questions follow: What do we mean by "ahead" and how large should any reconnaissance step be? The exercise of trying to answer these questions led to a procedure which may be visualized by studying Fig. 1, which depicts the situation for $n = 3$. We use this figure to introduce the reconnaisance procedure to the reader. The procedure consists of two phases: the first entails the generation of the reconnaissance points $\mathbf{x}^1$ and $\mathbf{x}^2$; and the second phase involves successive line searches along directions $\mathbf{s}^1$, $\mathbf{s}^2$ and $\mathbf{s}^3$.

Phase 1. At $\mathbf{x}^0$ take a step $\mathbf{x}^1 \leftarrow \mathbf{x}^0 + \alpha_0 \mathbf{p}^0$, where $\mathbf{p}^0 = -\mathbf{g}^0/\|\mathbf{g}^0\|$ and α_0 positive is chosen such that $f(\mathbf{x}^1) = f(\mathbf{x}^0)$. (This requirement, as will be shown later, need not be met precisely.) At $\mathbf{x}^1$ compute $\mathbf{g}^1$ and determine the normalized projection $\mathbf{p}^1$ of $\mathbf{g}^1$ orthogonal to $\mathbf{p}^0$. Take another step $\mathbf{x}^2 \leftarrow \mathbf{x}^1 + \alpha_1 \mathbf{p}^1$, again such that $f(\mathbf{x}^2) = f(\mathbf{x}^0)$, and compute $\mathbf{g}^2$. Finally, determine the projection of $\mathbf{g}^2$ orthogonal to both $\mathbf{p}^0$ and $\mathbf{p}^1$ to give $\mathbf{p}^2$. We have now generated reconnaissance points $\mathbf{x}^1$ and $\mathbf{x}^2$ on the level contours $\{\mathbf{x} | f(\mathbf{x}) = f(\mathbf{x}^0)\}$, as indicated in Fig. 1. The next phase involves the line searches.

Phase 2. A novel feature entering now is that we start our first line search from a point divorced from $\mathbf{x}^0$, namely from $\mathbf{y}^0 = \mathbf{x}^2$ along search direction $\mathbf{s}^1 = \mathbf{p}^2$. This yields the minimizing point $\mathbf{y}^1$. The next search direction $\mathbf{s}^2$ is taken as $\pm(\mathbf{x}^1 - \mathbf{y}^1)$, depending on which sign gives descent, with the minimizing point at $\mathbf{y}^2$. The last search direction is $\mathbf{s}^3 = \mathbf{x}^0 - \mathbf{y}^2$, giving the last minimizer $\mathbf{y}^3$ of the cycle. If the convergence criterion is not satisfied at $\mathbf{y}^3$ we may set $\mathbf{x}^0 \leftarrow \mathbf{y}^3$ and restart the cycle at Phase 1. □

We formalize the procedure for general n in the next section. Before doing so we discuss some of the features of the above outline that need further clarification. As pointed out, the requirement that $f(\mathbf{x}^i) = f(\mathbf{x}^0)$ is not essential. It seems reasonable however, that if we choose $f(\mathbf{x}^i) \lessapprox f(\mathbf{x}^0)$ we shall obtain sufficiently large steps to provide useful global information whilst, by ensuring that we remain within the level set $\{\mathbf{x} | f(\mathbf{x}) \leqslant f(\mathbf{x}^0)\}$ which contains $\mathbf{x}^*$, we prevent the optimal descent path from being distorted by using information too distant from $\mathbf{x}^*$. The precise procedure adopted to satisfy the specification that $f(\mathbf{x}^i) \lessapprox f(\mathbf{x}^0)$ is described in Section 6.

A further matter concerns the points from which the line searches are initiated. In the above we proposed that the searches be successively started from $\mathbf{y}^0$, $\mathbf{y}^1$ and $\mathbf{y}^2$. In the study of convergence

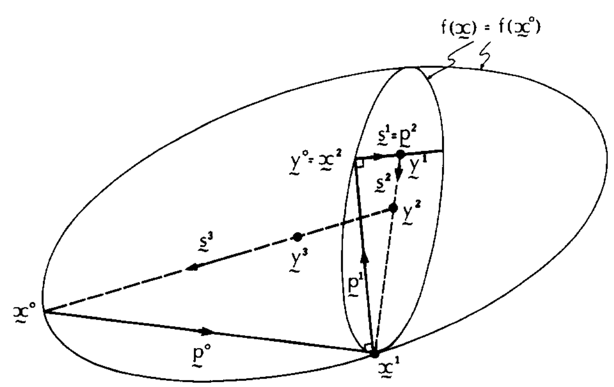

Fig. 1

that follows in Section 5 we show that global convergence is ensured if the last line search is initiated from $\mathbf{x}^0$ instead of $\mathbf{y}^2(\mathbf{y}^{n-1}$ in general). In practice, if the function is unimodal along the search direction the results will be identical and the search could thus be started from either $\mathbf{x}^0$ or $\mathbf{y}^{n-1}$. In fact, in the non-unimodel case one can conceive of situations where it may be preferable to initiate the last line search from $\mathbf{y}^{n-1}$. However, since global convergence cannot readily be demonstrated for this latter choice we will formally adhere to starting the last line search from $\mathbf{x}^0$.

3. THE FORMAL RECONNAISSANCE ALGORITHM

We now extend and formalize the above procedure to the general problem with n variables. The algorithm consists of a *basic cycle* which is repeated iteratively until convergence is achieved. In each cycle *two phases* are executed, the details of which can be stated as follows.

Phase 1: generation of reconnaissance points

1. Given $\mathbf{x}^0$, set $\mathbf{p}^0 \leftarrow -\mathbf{g}^0/\|\mathbf{g}^0\|$.
2. For $i = 0, 1, 2, \ldots, n-2$ set $\mathbf{x}^{i+1} \leftarrow \mathbf{x}^i + \alpha_i \mathbf{p}^i$, where the stepsize α_i is chosen such that $f(\mathbf{x}^{i+1}) \leqslant f(\mathbf{x}^0)$ (see Section 6) and where, for $i = 1, 2, \ldots, n-1$,

$$\mathbf{p}^i = \mathbf{q}^i/\|\mathbf{q}^i\|, \tag{1}$$

where

$$\mathbf{q}^i = -\mathbf{g}^i + \sum_{j=0}^{i-1} \beta_{j,i}\mathbf{p}^j \tag{2}$$

and

$$\beta_{j,i} = -\mathbf{g}^i \cdot \mathbf{p}^j$$

where $\cdot$ denotes the scalar product. (The expression for $\beta_{j,i}$ follows from the condition that $\mathbf{p}^i$ be orthogonal to $\mathbf{p}^j$, $j = 0, 1, 2, \ldots, i-1$.)

Phase 2: execution of line searches

1. Set $\mathbf{s}^1 \leftarrow \mathbf{p}^{n-1}$, $\mathbf{y}^0 \leftarrow \mathbf{x}^{n-1}$.
2. For $j = 1, 2, \ldots, n-1$ do:
   (i) set $\mathbf{y}^j \leftarrow \mathbf{y}^{j-1} + \lambda_j \mathbf{s}^j$, where λ_j minimizes $f(\mathbf{y}^{j-1} + \lambda \mathbf{s}^j)$ w.r.t. λ;
   (ii) set $\mathbf{s}^{j+1} \leftarrow \mathbf{r}^{j+1}/\|\mathbf{r}^{j+1}\|$, where $\mathbf{r}^{j+1} = \mathbf{x}^{n-j-1} - \mathbf{y}^j$;
   (iii) test whether convergence criterion is satisfied; if not continue, otherwise stop with $\mathbf{x}^* \doteq \mathbf{y}^j$.
3. Set $\mathbf{s}^n = -\mathbf{s}^n$ and set $\mathbf{y}^n \leftarrow \mathbf{x}^0 + \lambda_n \mathbf{s}^n$, where $\lambda_n > 0$ is the smallest value which yields a local minimum of $f(\mathbf{x}^0 + \lambda \mathbf{s}^n)$ w.r.t. λ.
4. Test whether convergence criterion is satisfied; if not set $\mathbf{x}^0 \leftarrow \mathbf{y}^n$ and start new cycle by going to Phase 1, otherwise stop with $\mathbf{x}^* \doteq \mathbf{y}^n$. □

4. QUADRATIC TERMINATION

We now show that the new method also possesses the desired property of quadratic termination, i.e. when applied to a positive-definite quadratic function

$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x} \cdot \mathbf{A}\mathbf{x} + \mathbf{b} \cdot \mathbf{x} + c$$

it terminates in, at most, n exact line searches. To demonstrate this we make use of the following well-known definition and related theorem.

Definition 4.1

Search directions $\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^n$ are *mutually conjugate* w.r.t. the positive-definite matrix $\mathbf{A}$ if $\mathbf{s}^i \cdot \mathbf{A}\mathbf{s}^j = 0$ for $i \neq j$. □

A method which generates such search directions is called a *conjugate direction method.*

Theorem 4.1 [2]

A conjugate direction method terminates for a quadratic function in, at most, n exact line searches and each $\mathbf{y}^i$ is the minimizer in the affine subspace generated by $\mathbf{y}^0$ and the directions $\mathbf{s}^1$, $\mathbf{s}^2, \ldots, \mathbf{s}^i$ (i.e. the set of points)

$$\left\{\mathbf{y} | \mathbf{y} = \mathbf{y}^0 + \sum_{j=1}^{i} \omega_j \mathbf{s}^j, \quad \forall\, \omega_j \in R\right\}. \qquad \square$$

Without loss in generality we assume in what follows that $\mathbf{p}^i \neq \mathbf{0}$, $i = 0, 1, 2, \ldots, n-1$. If $\mathbf{p}^m = \mathbf{0}$ for some m, then it simply means that we minimize in the affine subspace

$$D^m = \left\{\mathbf{x} | \mathbf{x} = \mathbf{x}^0 + \sum_{j=0}^{m-1} \gamma_j \mathbf{p}^j, \quad \forall\, \gamma_j \in R\right\}.$$

We now state and prove the following theorem relating to the reconnaissance method.

Theorem 4.2

The search directions $\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^n$ generated by the reconnaissance method are mutually conjugate w.r.t. the positive-definite matrix $\mathbf{A}$.

Proof. The proof is by induction. It can easily be shown that $\mathbf{s}^1$ and $\mathbf{s}^2$ are mutually conjugate. We give the general induction argument in which we assume that $\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^k$ are mutually conjugate and then prove that $\mathbf{s}^l \cdot \mathbf{A}\mathbf{s}^{k+1} = 0$ for $l = 1, 2, \ldots, k$. To do this we need the following expressions:

$$\mathbf{s}^1 = \mathbf{p}^{n-1}; \quad \mathbf{s}^i = \mathbf{x}^{n-i} - \mathbf{y}^{i-1}; \tag{3}$$

and from equation (2) in Section 3, we have

$$\mathbf{g}^i = \sum_{j=0}^{i} \beta_{j,i} \mathbf{p}^j, \tag{4}$$

with $\beta_{i,i} = -\|\mathbf{q}^i\|$, $i = 1, 2, \ldots, n$.

It can also be shown that the search directions are given by

$$\mathbf{s}^l = \sum_{j=1}^{l} \mu_j \mathbf{p}^{n-j}, \tag{5}$$

where the μs are dependent on and determined by the αs and λs. Also, since $\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^k$ are mutually conjugate they are linearly independent and together with $\mathbf{y}^0 = \mathbf{x}^{n-1}$ they generate the affine subspace

$$D^k = \left\{\mathbf{y} | \mathbf{y} = \mathbf{y}^0 + \sum_{j=1}^{k} \omega_j \mathbf{s}^j, \quad \forall\, \omega_j \in R\right\}.$$

Now, for $l = 1, 2, \ldots, k$,

$$\begin{aligned}
\mathbf{s}^l \cdot \mathbf{A}\mathbf{s}^{k+1} &= \mathbf{s}^l \cdot \mathbf{A}(\mathbf{x}^{n-k-1} - \mathbf{y}^k) \qquad \text{by equations (3)} \\
&= \mathbf{s}^l \cdot (\mathbf{A}\mathbf{x}^{n-k-1} + \mathbf{b}) - \mathbf{s}^l \cdot (\mathbf{A}\mathbf{y}^k + \mathbf{b}) \\
&= \mathbf{s}^l \cdot \mathbf{g}^{n-k-1} - \mathbf{s}^l \cdot \mathbf{g}(\mathbf{y}^k).
\end{aligned}$$

The latter term equals zero since $\mathbf{y}^k$ is a minimizer in D^k, and by equations (4) and (5) it follows that

$$\begin{aligned}
\mathbf{s}^l \cdot \mathbf{A}\mathbf{s}^{k+1} = \mathbf{s}^l \cdot \mathbf{g}^{n-k-1} &= \sum_{i=1}^{l} \mu_i \mathbf{p}^{n-i} \cdot \sum_{j=0}^{n-k-1} \beta_{n-k-1,j} \mathbf{p}^j. \\
&= 0 \text{ since the } \mathbf{p}\text{s are orthogonal.}
\end{aligned}$$

Thus, $\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^k$ and $\mathbf{s}^{k+1}$ are mutually conjugate w.r.t. $\mathbf{A}$, which completes the induction argument and thus the proof. $\square$

The above theorem together with Theorem 4.1 establishes the property of quadratic termination for the reconnaissance method.

5. GLOBAL CONVERGENCE

A simplified analysis of the global convergence property of the new algorithm may be carried out by observing that each cycle of the reconnaissance algorithm is completed by a final line search from $\mathbf{x}^0$ in the direction $\mathbf{s}^n$, as described in Steps 2(ii) and 3 of Phase 2 in Section 3. After this final line search the $\mathbf{x}^0$ for the next cycle is set $\mathbf{x}^0 \leftarrow \mathbf{y}^n$ or, if we wish to record the associated cycle numbers in parentheses we write: set

$$\mathbf{x}^0(k+1) \leftarrow \mathbf{y}^n(k) = \mathbf{x}^0(k) + \lambda_n(k)\mathbf{s}^n(k),$$

where $\lambda_n(k)$ minimizes $f(\mathbf{x}^0(k) + \lambda \mathbf{s}^n(k))$ w.r.t. λ.

Without loss in clarity we may now drop all superscripts and subscripts and write the general algorithm in simplified notation as follows.

Algorithm 5.1

For $k = 0, 1, 2, \ldots$ set

$$\mathbf{x}(k+1) \leftarrow \mathbf{x}(k) + \lambda(k)\mathbf{s}(k),$$

where $\lambda(k)$ minimizes $f(\mathbf{x}(k) + \lambda \mathbf{s}(k))$ w.r.t. λ, and $\mathbf{x}(0)$ denotes the overall starting point. □

We now define a descent method, according to the definition for a minimizing method given by Stoer and Bulirsch [3], and then make use of a further theorem of theirs to make a statement regarding the global convergence property of the reconnaissance algorithm.

Definition 5.1

If in a method of the form of Algorithm 5.1 we choose sequences $\sigma(k)$, $\lambda(k)$, $\gamma(k)$ and search directions $\mathbf{s}(k) \in D(\gamma(k), \mathbf{x}(k))$, where

$$D(\gamma, \mathbf{x}) = \{\mathbf{d} \in R^n | \mathbf{g}(\mathbf{x}) \cdot \mathbf{d} \leqslant -\gamma \|\mathbf{g}(\mathbf{x})\|, \quad \gamma > 0 \quad \text{and} \quad \|\mathbf{d}\| = 1\},$$

and

$$\lambda(k) \in [0, \sigma(k) \|\mathbf{g}(\mathbf{x}(k))\|] = I_k$$

is the minimizer w.r.t. $\lambda \in I_k$, then the algorithm is called a *descent method.* □

The required global convergence theorem of Stoer and Burlirsch [3] now follows.

Theorem 5.1

Suppose that the level set $\{\mathbf{x} | f(\mathbf{x}) \leqslant f(\mathbf{x}(0))\}$ is compact. Then, if

$$1 \geqslant \sup_k \gamma(k) \geqslant \inf_k \gamma(k) > 0 \quad \text{and} \quad \inf_k \sigma(k) > 0$$

and if $f \in C^1$, the sequence $\mathbf{x}(k)$ of the associated descent method is well-defined and has at least one limit point where the gradient is zero. □

If we can now show that for the reconnaissance method, as represented by Algorithm 5.1, $\gamma(k)$ and $\sigma(k)$ satisfy the conditions of Theorem 5.1 then we have global convergence for the new method as well. To do this we require a further lemma.

Lemma 5.1

Suppose that the set $K \subset R^n$ is compact. Then for an arbitrary pair of points $\mathbf{x}, \mathbf{y} \in K$ such that $\|\mathbf{y} - \mathbf{x}\| = \alpha > 0$, there exists a positive constant γ, depending on α, such that

$$\frac{(\mathbf{y}-\mathbf{x})}{\|\mathbf{y}-\mathbf{x}\|} \cdot \frac{(\mathbf{z}-\mathbf{x})}{\|\mathbf{z}-\mathbf{x}\|} \geqslant \gamma > 0 \quad \text{and} \quad \gamma \leqslant 1$$

for every $\mathbf{z} \in W = \{\mathbf{w} | (\mathbf{w}-\mathbf{y}) \cdot (\mathbf{y}-\mathbf{x}) = 0, \ \mathbf{w} \in K\}$.

Proof. Select an arbitrary $\mathbf{z} \in W$. Since

$$(\mathbf{z}-\mathbf{y})\cdot(\mathbf{y}-\mathbf{x})=0,$$

it follows that

$$(\mathbf{z}-\mathbf{x}+\mathbf{x}-\mathbf{y})\cdot(\mathbf{y}-\mathbf{x})=0,$$

and thus

$$(\mathbf{z}-\mathbf{x})\cdot(\mathbf{y}-\mathbf{x})=\|\mathbf{y}-\mathbf{x}\|^2=\alpha^2$$

and consequently

$$\frac{(\mathbf{z}-\mathbf{x})}{\|\mathbf{z}-\mathbf{x}\|}\cdot\frac{(\mathbf{y}-\mathbf{x})}{\|\mathbf{y}-\mathbf{x}\|}=\frac{\alpha}{\|\mathbf{z}-\mathbf{x}\|}\geqslant\frac{\alpha}{\sup\limits_{\mathbf{x}',\mathbf{y}'\in K}\|\mathbf{x}'-\mathbf{y}'\|}=\gamma>0$$

and obviously $\gamma \leqslant 1$. □

The global convergence theorem now follows.

Theorem 5.2

Suppose f is continuously differentiable and the level set $K=\{\mathbf{x}|f(\mathbf{x})\leqslant f(\mathbf{x}(0))\}$ is compact and that all the reconnaissance points are contained in K. Then the reconnaissance method is convergent to a stationary point of f if

$$\inf_k \alpha_0(k)=\alpha_0>0.$$

Proof. With reference to Lemma 5.1 we note the equivalence

$$\mathbf{x}=\mathbf{x}(k)=\mathbf{x}^0(k),$$
$$\mathbf{y}=\mathbf{x}(k)-\alpha_0(k)\mathbf{g}(\mathbf{x}(k))/\|\mathbf{g}(\mathbf{x}(k))\|,$$
$$\mathbf{z}=\mathbf{y}^{n-1}(k).$$

Thus, applying Lemma 5.1 we have, since

$$\inf_k \alpha_0(k)\geqslant\alpha_0>0,$$

that

$$\frac{-\mathbf{g}(\mathbf{x}(k))}{\|\mathbf{g}(\mathbf{x}(k))\|}\cdot\mathbf{s}(k)\geqslant\frac{\alpha_0(k)}{\sup\limits_{\mathbf{x}',\mathbf{y}'\in K}\|\mathbf{x}'-\mathbf{y}'\|}=\gamma(k)>0$$

and $\gamma(k)\leqslant 1$, $\forall k$, and obviously also

$$\inf_k \gamma(k)>0.$$

Therefore, provided we choose $\sigma(k)$ equal to a sufficiently large positive constant for all cycles, the conditions of Theorem 5.1 are satisfied and we thus have convergence to a stationary point. □

Of course the condition $\alpha_0>0$ is sufficient and may not be necessary. In practice the above theorem means that if, in the execution of the algorithm, we set a lower bound α_0 on $\alpha_0(k)$, whilst ensuring that the reconnaissance points remain in K, we will have convergence to a stationary point. Although it seems likely that this will usually be the case for sufficiently smooth f, it does not altogether exclude the possibility that as the algorithm progresses it may not be possible to generate reconnaissance points in K if a lower bound is set on $\alpha_0(k)$.

The requirement that

$$\inf_k \alpha_0(k) > 0$$

can be relaxed by a somewhat more subtle argument. We introduce the notation

$$\epsilon_k = f(\mathbf{x}(k)) - f(\mathbf{x}(k) + \alpha_0(k)\mathbf{p}^0(k))$$

for the final theorem.

Theorem 5.3

Suppose that f is twice continuously differentiable and the level set K is compact and that all the generated reconnaissance points are contained in K. Then, if we can choose $\alpha_0(k) > 0$ such that

$$\epsilon_k \leqslant \tfrac{1}{2}\alpha_0(k) \| \mathbf{g}(\mathbf{x}(0)) \|, \tag{6}$$

the algorithm is convergent to a stationary point.

Proof. Since K is bounded condition (6) can always be satisfied provided $\mathbf{g}(\mathbf{x}(0)) \neq \mathbf{0}$, and since $f \in C^2$ and K is compact, it follows that there exists a finite η such that

$$\begin{aligned} f(\mathbf{x}^0(k)) = f(\mathbf{x}^1(k)) + \epsilon_k \leqslant f(\mathbf{x}^0(k)) \\ + \alpha_0(k)\mathbf{g}(\mathbf{x}^0(k))\cdot\mathbf{p}^0(k) + \tfrac{1}{2}\eta(\alpha_0(k))^2 + \epsilon_k \\ \leqslant f(\mathbf{x}^0(k)) - \tfrac{1}{2}\alpha_0(k)\|\mathbf{g}(\mathbf{x}^0(k))\| + \tfrac{1}{2}\eta(\alpha(k))^2, \end{aligned}$$

by the definition of $\mathbf{p}^0$ and condition (6), resulting in the bound

$$\|\mathbf{g}(\mathbf{x}^0(k))\| \leqslant \eta\alpha_0(k). \tag{7}$$

Now either $\inf \alpha_0(k) = 0$ or $\inf \alpha_0(k) > 0$. In the first case, bound (7) implies convergence to a stationary point. On the other hand, if the second possibility occurs, it follows from the previous Theorem 5.2 that the method yields a stationary point.

6. PRACTICAL GENERATION OF RECONNAISSANCE POINTS

Step 2 of Phase 1 (Section 3) requires the generation of reconnaissance points such that

$$f(\mathbf{x}^{i+1}) = f(\mathbf{x}^i + \alpha_i\mathbf{p}^i) \lessapprox f(\mathbf{x}^0),$$

for $i = 0, 1, 2, \ldots, n-2$. We now describe how this may economically be done in practice. Note that the determination of a suitable α_i is essentially a one-dimensional problem: find a sufficiently large α $(=\alpha_i)$ such that

$$F(\alpha) \lessapprox F(0), \tag{8}$$

where $F(\alpha) = f(\mathbf{x}^i + \alpha\,\mathbf{p}^i)$, $F'(0) < 0$ and F is assumed to be continuously differentiable. We now attempt to satisfy the above requirement via the related supplementary problem: given a constant $\rho \in (-\infty, 0]$ find an $\alpha > 0$ such that

$$F(\alpha) \leqslant F(0) + \rho\alpha F'(0). \tag{9}$$

If a solution to this problem exists it may be found by performing the following algorithm.

Algorithm 6.1

1. Find a trial steplength $\bar{\alpha}$.
2. Find the smallest integer $m > 0$ such that for a specified $\omega \in (0, 1]$

$$F(l_m) \geqslant F(0) + \omega l_m F'(0),$$

where

$$l_m = \bar{\alpha} \sum_{i=1}^{m} i!$$

3. Find the smallest integer $n > 0$ such that

$$F(\alpha_{nm}) \leqslant F(0) + \rho\alpha_{nm}F'(0),$$

where $\alpha_{nm} = l_m - \bar{\alpha}m!(1 - \delta^{n-1})$ and $\delta \in (0, 1)$ is specified by the user. If a solution to problem (9) exists we take $\alpha = \alpha_{nm}$.

The original requirement (8) forces the choice $\rho = 0$ and for $F(\alpha)$ to be "close" to $F(0)$, ω should be close to zero and δ near the value 1. For economy, however, we typically choose $\omega = 0.5$ and $\delta = 0.5$, which in practice yields sufficiently large reconnaissance steps. We expand regarding the choice of $\bar{\alpha}$. The procedure should be such that if the function is quadratic it yields an $\bar{\alpha}$ such that $F(\bar{\alpha}) = F(0)$. This may be done by initially fitting a quadratic to the points $F(0)$, $F(\epsilon)$ and $F(2\epsilon)$, where the parameter ϵ is chosen sufficiently small such that the resulting quadratic function $P(\alpha)$ is a local approximation of F near 0 and such that a sufficiently accurate finite difference approximation to $F'(0)$ may be calculated. Typically, we choose $\epsilon = 10^{-6}$. We now determine the value $\bar{\alpha} \neq 0$ such that $P(\bar{\alpha}) = F(0)$.

Two problems may now arise. Firstly, if $\bar{\alpha} < 0$ (i.e. if P has a maximum). In this case we set $\bar{\alpha} \leftarrow l$, some prescribed lower bound for $\bar{\alpha}$, and continue performing Steps 2 and 3 of the above algorithm. Secondly, the calculated value for $\bar{\alpha}$ may be either exceedingly small or excessively large. Here we set

$$\bar{\alpha} \leftarrow l \quad \text{if} \quad \bar{\alpha} < l$$

and

$$\bar{\alpha} \leftarrow u \quad \text{if} \quad \bar{\alpha} > u,$$

where u is some prescribed upperbound for $\bar{\alpha}$.

Finally, in practice we may avoid the calculation of the α_is for each cycle by adopting the following heuristic. Compute the α_is every qth cycle, i.e. for $k = nq$, $n = 0, 1, 2, \ldots,$ and calculate an average steplength

$$\alpha_a = \sum_{i=1}^{n-1} \alpha_i(k)/(n-1).$$

For the intermediate cycles we then use

$$\alpha_i(l) = \alpha_a \|\mathbf{g}(\mathbf{x}^0(l))\|^2/\|\mathbf{g}(\mathbf{x}^0(k))\|^2, \qquad i = 0, 1, 2, \ldots, n-2, \tag{10}$$

for $l = k+1, k+2, \ldots, k+q-1$.

7. AN EXAMPLE

The main purpose of this paper has been the presentation and analysis of the new method rather than an experimental study of its performance relative to others. In conclusion it is nevertheless appropriate that we illustrate the working of the new algorithm by presenting the computed path for the well-known Rosenbrock two-dimensional test function

$$f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

with starting point $\mathbf{x}_0 = (-1.2, 1)^T$. This function has a steep-sided valley which curves along $x_2 = x_1^2$ with the minimum occurring at the base of the valley at $\mathbf{x}^* = (0, 0)^T$ with $f(\mathbf{x}^*) = 0$.

The solid circles in Fig. 2 indicate the path for the reconnaissance method. Each circle denotes the resultant point after each cycle (two line searches). For the given path the α_is for the reconnaissance points were computed for each cycle according to Algorithm 6.1 outlined in the previous section. The reconnaissance method progresses in large steps along the valley yielding $\mathbf{x}(8) = (1.0000000, 1.0000000)^T$ to eight significant digits with $f(\mathbf{x}(8)) = 10^{-16}$. The required line searches were carried out using the IMSL line search subroutine ZXLSF with an absolute accuracy of 10^{-6}. Calculating the α_is only every third cycle and using equation (10) for the intermediate cycles gives an almost identical path whilst cutting the total number of function evaluations required for the generation of the reconnaissance points to the relatively insignificant number of 16. For

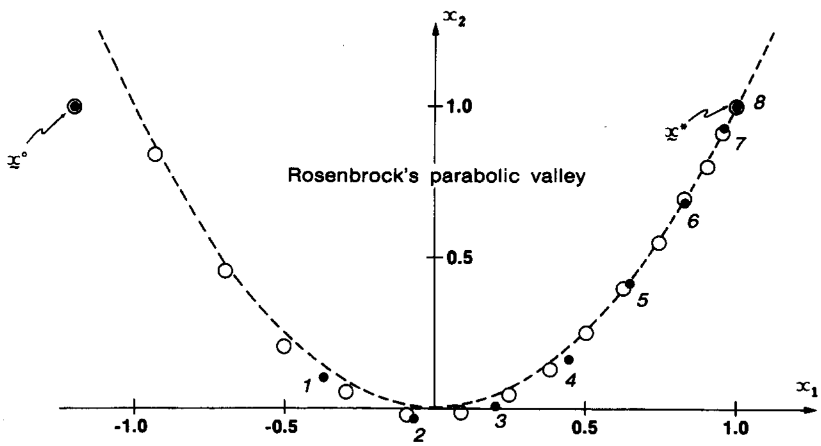

Fig. 2

comparison we also give the path generated by the traditional Fletcher–Reeves method [1]. The open circles denote the position before each reset (also after two line searches). The path is similar to that of the reconnaissance method although requiring almost twice the number of cycles.

8. CONCLUSION

The method presented here has certain desirable features. It does not require any explicit information regarding second derivatives. The method possesses the property of quadratic termination and it is shown that for the general non-quadratic case it is globally convergent under comparatively mild conditions. The indications are that the required reconnaissance points may be generated relatively inexpensively with few function evaluations required in addition to the gradient evaluation per reconnaissance point. It is also expected that the use of reconnaissance points allows for rapid progress for functions that possess steep and curved valleys leading to the minima.

A drawback and point of criticism, compared with the Fletcher–Reeves method, is that the new method requires the storage of two sets of n-vectors per cycle, namely the reconnaissance points $\mathbf{x}^i$ and the normalized gradients $\mathbf{p}^i$, $i = 0, 1, \ldots, n-1$. Also we have assumed, and in the example attempted to execute, exact line searches that in practice would make the method uneconomical. It is of course possible to use inexact line searches by applying safe-guarded parabolic or cubic approximations to the behaviour of the function together with some rule, such as the Goldstein–Armijo principle [4], to ensure sufficient descent and resultant convergence. With such modifications one may expect the method to be competitive from the point of the number of overall function evaluations. A detailed experimental study of the reconnaissance method with inexact line searches and a comparison of its performance with other competitive methods remains to be done.

REFERENCES

1. R. Fletcher and C. M. Reeves, Function minimization by conjugate gradients. *Comput. J.* **7**, 149–154 (1964).
2. R. Fletcher, *Practical Methods of Optimization.* Vol. 1: *Unconstrained Optimization*, p. 24. Wiley, New York (1980).
3. J. Stoer and R. Burlirsch, *Introduction to Numerical Analysis*, p. 258. Springer-Verlag, Berlin (1980).
4. P. E. Gill, W. Murray and M. H. Wright, *Practical Optimization*, p. 100. Academic Press, New York (1981).